COGNATE BEGINNINGS TO BILINGUAL LEXICAL ACQUISITION[*]

A PREPRINT

Gonzalo Garcia-Castro
Center for Brain and Cognition
Universitat Pompeu Fabra
Barcelona, 08005
gonzalo.garciadecastro@upf.edu

Daniela S. Avila-Varela
Center for Brain and Cognition
Universitat Pompeu Fabra
Barcelona, 08005
avila.varela.daniela@gmail.com

Ignacio Castillejo
Departamento de Psicología
Universidad Autónoma de Madrid
Madrid, 28049
jignaciocastillejo@gmail.com

Nuria Sebastian-Galles
Center for Brain and Cognition
Universitat Pompeu Fabra
Barcelona, 08005
nuria.sebastian@upf.edu

August 12, 2023

ABSTRACT

Bilingual infants' developmental trajectories of lexical acquisition are equivalent to their monolingual peers'. This is remarkable, given the complexity of their linguistic input. Recent studies suggest that bilingual vocabulary growth is boosted by the number of cognates (form-similar translation equivalents) shared by the pair of languages being learned, and that this cognateness facilitation effect is driven by a stronger parallel activation of cognates during linguistic exposure, compared to non-cognates. The mechanisms behind this facilitation are still unclear. In this study, we propose an account of bilingual lexical acquisition in which parallel activation increases the rate at which children accumulate learning instances for words in both languages, even in fully monolingual situations. We predicted a stronger cognate facilitation for words to which children were exposed less frequently (low-exposure words), as they are co-activated by their translation more often than high-exposure words. We developed an extensive online vocabulary checklist, the Barcelona Vocabulary questionnaire (BVQ), to collect vocabulary data from 366 Catalan-Spanish bilingual toddlers aged 12 to 32 months. We used Bayesian explanatory item response theory to model the acquisition trajectories of 604 Catalan and Spanish words. We found an interaction between exposure and cognateness, suggesting that cognateness facilitates the aquisition of low-exposure words, but not of mean exposure or high-exposure words. Overall, our findings suggest that cognateness plays a key role in bilingual lexical acquisition, and provides evidence for a frequency-mediated facilitation effect driven by parallel activation.

Keywords cognate • word acquisition • vocabulary • bilingualism • item response theory • bayesian

[*]The authors declare no conflicts of interest with regard to the funding source of this study. This study was supported by the Spanish Ministry for Science and Innovation and State Research Agency (Project PID2021- 123416NB-I00 financed by MCIN/ AEI/ 10.13039/501100011033 / FEDER, UE) and the Economic and Social Research Council (ESRC) (ES/S010947/1, UK). GGC was supported by a FPI research contract (PRE2019-088165). DAV was supported by the European Union's Horizon 2023 research and innovation program under Marie Skodowska–Curie Grant (765556) and a postdoctoral fellowship from the Foundation for Science and Technology of Portugal (UIDB/00214/2020). IC was supported by the Investigo program funded by the European Union's NextGenerationEU (NGEU) recovery plan. NSG was supported by an ICREA Academia award from the Catalan Institution for Research and Advanced Studies (ICREA). We are grateful to Chiara Santolin, Ege E. Özer, and the rest of the Speech Acquisition and Perception research group, and to Alicia Franco-Martínez and Cristina Rodríguez-Prada, for their helpful feedback. We thank Xavier Mayoral, Silvia Blanch, and Cristina Cuadrado for their technical support, and Cristina Dominguez and Katia Pistrin for their efforts in recruiting infants. We also thank all families and infants who participated in the experiments.

1 Introduction

The foundations of word learning are in place at an early age. At six months, infants start directing their gaze to objects when hearing their labels (Bergelson & Swingley, 2012, 2015; Tincoff & Jusczyk, 1999), and shortly after caregivers start reporting some words as acquired by their infant in vocabulary checklists (e.g., Fenson et al., 2007; Samuelson, 2021). Most research on early word acquisition relies extensively on data from monolingual children, and is oblivious to the fact that a substantial proportion of the world population acquires more than one language from early ages (Grosjean, 2021). Previous work on bilingual vocabulary acquisition pointed to bilingual toddlers knowing, on average, less words in each of their languages than their monolinguals peers, and to both groups knowing a similar number of words—if not more words—when the bilinguals' two languages are pooled together. Hoff et al. (2012) found that English-Spanish bilingual toddlers in South Florida (United States) knew less words in English than monolinguals did, but both groups knew a similar total amount of words when both English and Spanish vocabularies were counted together. Other studies have provided converging evidence that bilinguals know a similar or even larger number of words than monolinguals when the two languages are aggregated (Gonzalez-Barrero et al., 2020; Oller & Eilers, 2002; Patterson, 2004; Patterson & Pearson, 2004; Pearson & Fernández, 1994; Smithson et al., 2014). A more detailed analysis of the words in bilinguals' lexicons shows some interesting patterns.

One important observation of studies on bilinguals' early vocabulary acquisition is that cognate words are easier to acquire than non-cognate words. Cognate words are translation equivalents that are phonologically similar (or share some type of form-similarity). For instance, the Spanish translation equivalent of *cat* is *gato*, a cognate word; the translation equivalent of *dog* is *perro*, a non-cognate word. The differences in the percentages of cognate and non-cognate words between two languages is related to historical reasons: languages typologically close (like Dutch and English or Italian and Spanish) share more cognates than languages typologically distant (like English and Chinese, or Urdu and Spanish). The conclusion that cognate words are easier to learn is based on two types of evidence: studies investigating vocabulary sizes in children learning language pairs with different percentages of cognates (that is, differing in their typological distance) and studies comparing the number of cognate and non-cognate words children know in a specific language pair.

Floccia et al. (2018) published an impressive study comparing vocabularies of children learning several language pairs differing in their percentage of cognates. The authors collected vocabulary data on word comprehension and production from 372 24-month-old bilingual toddlers living in the United Kingdom who were learning English and an additional language. The additional language was one of 13 typologically diverse languages: Bengali, Cantonese Chinese, Dutch, French, German, Greek, Hindi or Urdu, Italian, Mandarin Chinese, Polish, Portuguese, Spanish, and Welsh. The authors calculated the average lexical overlap between the words in each of these additional languages and their translation equivalents in English. Lexical overlap was calculated in terms of phonological similarity (described below) and it was taken as a proxy of the degree of cognateness between each pair of languages. Floccia and co-workers reported an increase in vocabulary size in the additional language (i.e., not English) associated with an increase in the average phonological similarity between the translation equivalents of each language pair. For example, English-Dutch bilinguals (languages with a high phonological overlap), were able to produce more Dutch words than English-Mandarin bilinguals (languages with a low phonological overlap) were able to produce in Mandarin. Blom et al. (2020), Bosma et al. (2019), and Gampe et al. (2021) reported similar results, providing converging evidence of a facilitatory effect of a lower language distance (i.e., higher degree of cognateness) on vocabulary size.

A second set of studies suggested that cognates are overrepresented in bilinguals' early lexicon. Bosch & Ramon-Casas (2014) collected parental reports of expressive vocabulary from 48 Catalan-Spanish bilinguals aged 18 months and found that cognates represented a larger proportion of vocabulary than non-cognates. Schelletter (2002) provided converging evidence from a longitudinal single-case study, in which an English-German bilingual child produced cognates earlier than non-cognates, on average. But the high proportion of cognates in the vocabulary of the participants in these two studies may not necessarily evidence of a facilitation effect of cognateness, but rather of simply the high proportion of cognates present in the pair of languages being learned. For instance, if two given languages share a high proportion of cognates like 70%, the vocabulary contents of children learning both languages should, in principle, approximate such proportion of cognates, even in the absence of a cognateness facilitation effect. More recently, Mitchell et al. (2022) addressed this issue in a longitudinal study. The authors collected expressive vocabulary data of 47 16- to 30-month-old French-English bilinguals living in Canada, in both languages. They created two lists of translation equivalents: one made of 131 cognates, and one made of 406 non-cognates. The proportion of words that children were reportedly able to produce was higher in the cognate lists than in the non-cognate list across ages. Critically, this difference persisted after both lists were matched in size, controlling their semantic category (i.e., furniture, animals, food were similarity represented in both lists) and age-of-acquisition norms (an index of word difficulty). Taken together, the results of these two lines of research have been considered as supporting to the hypothesis that phonological similarity (as reflected in cognateness) plays a facilitation role in bilingual word acquisition.

Parallel activation of bilinguals' lexicons has been proposed as the underlying mechanism for such facilitatory effect (e.g., Floccia et al., 2018; Mitchell et al., 2022). The parallel activation hypothesis stems from the language non-selective account of lexical access, which suggests that bilinguals activate both languages simultaneously during language processing, even in fully monolingual contexts. Evidence with adult bilinguals supporting the language-non selective account of lexical access has been reported for language comprehension and production, across the auditory and visual (reading and signing) modalities (Gimeno-Martínez et al., 2021; Hoshino & Kroll, 2008; Morford et al., 2011; Shook & Marian, 2012; Spivey & Marian, 1999; see Kroll & Ma, 2017 for review). One of the clearest pieces of evidence of parallel activation was provided by Costa et al. (2000). In this study, Spanish monolinguals and Catalan-Spanish bilingual adults were asked to name pictures of common objects in Spanish. In half of the trials, the object labels were cognates in Spanish and Catalan (*árbol-arbre*, translations of *tree*), whereas in the other half of the trial labels were non-cognates (*mesa-taula*, translations of *table*), obviously, such distinction was only relevant for bilinguals. Bilinguals named cognate pictures faster than non-cognate pictures, even after adjusting for the lexical frequency of the items. In contrast, Spanish monolinguals, who were unfamiliar with the Catalan translations of the Spanish words they uttered, showed equivalent naming times for the two types of stimuli. The authors interpreted the difference between cognates and non-cognates in bilinguals as reflecting the additional phonological activation that cognate words would receive from their translation equivalents (due to language non-selective activation of bilinguals' lexicons). These results showed that bilinguals' Catalan phonology was activated during the production of Spanish words, facilitating the naming of cognate pictures. Evidence of parallel activation has been reported in bilingual toddlers and children (Bosma & Nota, 2020; Floccia et al., 2020; Poarch & Hell, 2012; Poulin-Dubois et al., 2013; Von Holzen et al., 2019; Von Holzen & Mani, 2012).

Although there is a consensus on the role of parallel activation in bilinguals' lexical processing and acquisition, previous studies do not address its influence on the learning trajectories of words. Results are aggregated across words and provide no information about the specific dynamics of how parallel activation influences word learning. This is the goal of the present research.

We propose an account in which a learning instance for a word may also represent a learning instance for its translation equivalent, to the extent that such translation equivalent is co-activated. We use the term *learning instance* in the fashion of accumulator models of language acquisition; as an exposure to a word-form that constitutes an opportunity for the child to accumulate information about the word. We do not assume if a learning instance is a discrete or a continuous unit of accumulation of information. We consider that a learning instance of a word consists of an exposure to the its phonological form if the resulting strength of activation of its phonological representation in the child's lexicon reaches some thoretical threshold that leads to word-form recognition. This activation may result from the infant being exposed to the actual word-form, or the result of activation spreading through phonological or semantic links across lexical representation, as in the case of parallel activation. The strength of this co-activation is proportional to the phonological similarity between the two translation equivalents; given that cognates share higher phonological similarity than non-cognates, the former should be co-activated more strongly than the latter. This should lead to a faster accumulation of learning instances for cognates, compared to non-cognates. Parallel activation would allow bilingual children to accumulate learning instances for words in both languages even during fully monolingual situations, but the impact of this mechanism would be asymmetric across languages: words from the lower-exposure language would receive stronger activation from words in the higher-exposure language than vice versa. Therefore, the acquisition of words from the lower-exposure language would benefit more strongly from their cognate status than words from the higher-exposure language. This asymmetric cross-language activation would be consistent with previous reports of larger priming effects from the dominant to the non-dominant language.

Consider the example of the Catalan-Spanish cognate translation equivalent /ˈgat/–/ˈga.to/ [*cat*], which are phonologically very similar. When the child listens to /ˈgat/, they will strongly co-activate /ˈga.to/ in parallel. If the child has already formed a form-meaning association for both word-forms, parallel activation may result from the activation of their common concept or from activation spreading throught phonological similarity. We assume semantic co-activation to be constant across cognate and non-cognate translation equivalents, and focus on phonological co-activation as an additional source of activation that affects cognates more strongly than non-cognates. Therefore, this exposure will count as a learning instance for both co-activated forms. The case of the non-cognate translation equivalent /ˈgos-ˈpe.ro/ [*dog*] would be different. Given the low phonological similarity between both word-forms, an exposure to ˈgos will result in a weak activation of /ˈpe.ro/ leading to such exposure counting as a learning instance for /ˈgos/ (which the child was exposed to), but not for /ˈpe.ro/. While /ˈgat/–/ˈga.to/ will benefit from phonological co-activation, /ˈgos-ˈpe.ro/ will not. If the child receives linguistic input from one of the languages more often than from the other, this effect might affect each form of the cognate translation equivalent differently. For instance, if the child receives a larger amount of Catalan input than Spanish input, they will encounter the Catalan form /ˈgat/ more

Table 1: Summary of the items included in the final analyses.

Semantic category	List A	List B	List C	List D	Examples
Household items	31	26	30	25	clock, video
Food and drink	29	26	23	27	sausage, yogurt
Animals	26	23	19	25	panther, tiger
Outside	14	13	13	15	farm, stone
Body parts	14	12	11	11	face, finger
Toys	11	11	12	13	piano, racket
Clothes	12	12	10	10	zipper, sandal
Vehicles	9	10	11	10	helicopter, tractor
Colours	6	6	6	6	red, green
People	7	4	6	6	police, babysitter
Furniture and rooms	4	4	4	4	corridor, terrace
Time	2	2	2	2	day, night
Adventures	1	1	1	1	witch
Parts of things	1	1	1	1	wheel
N	167	151	149	156	-

frequently than the Spanish form /ˈga.to/. Through parallel activation, /ˈgat/ will activate /ˈga.to/ more often than vice versa. Ultimately, /ˈga.to/ will benefit more strongly from its cognate status than /ˈgat/, as it receives additional learning instances from its translation equivalent more often than /ˈgat/.

To test these predictions, we collected vocabulary data on production and comprehension from a large sample of bilingual Catalan-Spanish children. We adopted a Bayesian explanatory item response theory approach to model the probability of acquisition of 604 Catalan and Spanish nouns included in the vocabulary checklist. Words were considered as acquired if caregivers reported such word to be understood (comprehension) or understood and said (production) by their child. We estimated the impact of several predictors of interest on the probability of acquisition, including participants' age and rate of exposure to the word-form, and the cognate status of the word-form. As described in the methods section, rate of exposure was a composite measure taking into account participant' language exposure and word's lexical frequency. We predicted an interaction between cognate status and word-form exposure rate in which the probability of comprehension is higher for low-exposure cognate words, but not for high-exposure cognate words.

2 Methods

All materials, data, and reproducible code can be found at the OSF (https://osf.io/hy984/) and GitHub (https://github.com/gongcastro/cognate-beginnings) repositories. This study was conducted according to guidelines laid down in the Declaration of Helsinki, and was approved by the Drug Research Ethical Committee (CEIm) of the IMIM Parc de Salut Mar, reference 2020/9080/I.

2.1 Questionnaire

To collect vocabulary data from participants, we created an *ad hoc* questionnaire: the Barcelona Vocabulary Questionnaire (BVQ) (Garcia-Castro et al., 2023). This questionnaire was inspired by the MacArthur-Bates Communicative Development Inventory (Fenson et al., 2007) and its adaptations to other languages, and was implemented on-line using the formr platform (Arslan et al., 2020). This questionnaire is structured in three blocks: (1) a language exposure questionnaire, (2) a demographic survey, and (3) two vocabulary checklists. Vocabulary checklists followed a similar structure as the Oxford Communicative Developmental Inventory (OCDI) (Hamilton et al., 2000) and consisted in two lists of words: one in Catalan and one in Spanish. Both lists included items from a diverse sample of 26 semantic or functional categories. The Catalan checklist contained 793 items and the Spanish checklist contained 797. Items in one language were translation equivalents of the items in the other language (e.g., the same participant responded to both *gos* and *perro*, Catalan and Spanish for *dog*), roughly following a one-to-one mapping. Some of the words in Catalan did not have a clear translation or had more than one possible translation in Spanish, and vice versa, therefore the unequal number of words included in the two lists (see Table 1 for a summary of the questionnaire items).

Table 2: Participant sample size by age and degree of exposure to Catalan.

	Age (months)					
Catalan exposure	[10-14]	(14, 18]	(18, 22]	(22-26]	(26-30]	(30-34]
75-100%	18	23	37	38	20	7
50-75%	8	13	30	41	18	1
25-50%	10	18	46	30	17	0
0-25%	7	11	21	17	8	1
N	43	65	134	126	63	9

For each word included in the vocabulary checklists, we asked parents to report whether their child was able to understand it, understand *and* say it, or did not understand or say it (checked out by default). Given the large number of words in the vocabulary checklists, we created four different subsets of the complete list of items. Each subset contained a random but representative sub-sample of the items from the complete list (see Table 1). Semantic or functional categories with less than 16 items—thus resulting in less than four items after dividing it in four lists—were not divided in the short version of the questionnaire: all of their items were included in the four lists. Items that were part of the trial lists of some ongoing experiments in the lab were also included in all versions. The resulting reduced list contained between 343 and 349 Catalan words, and between 349 and 371 Spanish words. Participants were randomly allocated into one of the four subsets.

To compute predictors of interest, we manually generated a broad phonological transcription of every word included in the vocabulary checklists in X-SAMPA format (Wells, 1995). Catalan word-forms were transcribed to Central Catalan phonology, and Spanish word-forms were transcribed to Castilian Spanish phonology.

2.2 Participants

We collected 440 responses to the questionnaire from 369 distinct children from the Metropolitan Area of Barcelona between the 30th of marzo, 2020 and the 31th of octubre, 2022: 314 of those participants participated once, 43 twice, 8 three times, and 4 four times. Recurrent participants provided responses with a minimum of 25 days between responses, and a maximum of 527, and were always allocated to the same questionnaire list (A, B, C, or D). Participants were part of the database of the Laboratori de Recerca en Infància (Universitat Pompeu Fabra) and were contacted by e-mail or phone if their child was aged between 12 and 32 months, and had not been reported to be exposed more than 10% of the time to a language other than Spanish or Catalan (see Table 2 for a more detailed description of the sample). In total, 70 participants (15.91%) participants were reported to be exposed to a third language other than Catalan and Spanish. All families provided informed consent before participating. Upon consent, families were sent a link to the questionnaire via e-mail, which they filled from a computer, laptop, or mobile device. Filling the questionnaire took 30 minutes approximately. After completion, families were rewarded with a token of appreciation.

We used the highest self-reported educational attainment of parents or caregivers as a proxy of participants' socio-economic status (SES). This information was provided by each parent or caregiver by selecting one of six possible alternatives in line with the current educational system in Spain: *sense escolaritzar/sin escolarizar* [no education], *educació primària/educación primaria* [primary school], *educació secundària/educación secundaria* [secondary school], *batxillerat/bachillerato* [complementary studies/high school], *cicles formatius/ciclos formativos* [vocational training], and *educació universitària/educación universitaria* [university degree]. Most families reported university studies (358, 82%), followed by families were the highest educational attainment were vocational studies (61, 14%), secondary education (8, 2%), complementary studies (6, 1%), primary education (1, <1%), and no formal education (2, <1%).

2.3 Data analysis

2.3.1 Data processing

We collected data for 1,590 words. We restricted the analyses to responses to nouns (628 items corresponding to other grammatical classes were excluded, see Fourtassi et al., 2020 for a similar approach). We then excluded items with missing lexical frequency scores (n = 269, see Section 2.3.3), items that included more than one lemma (e.g., *mono/mico* [monkey], n = 48), multi-word items or phrases (e.g., *barrita de cereales* [cereal bar], n = 9). Finally, we removed items without a translation in the other language (n = 32). This resulted in a final list of 604 items, corresponding to 302 Catalan words and their 302 Spanish translations (302 translation equivalents). After collecting participants' responses, the final dataset consisted of 139,326 observations, each corresponding to a single response of one participant to one item. Each translation equivalent received a median of 234 responses (*Min* = 108, *Max* = 880) from participants, both languages pooled together. Data processing and visualisation was done in R (R Core Team, 2013, version 4.2.2) using the *tidyverse* family of packages (Wickham et al., 2019).

2.3.2 Modelling approach

We modelled the probability of participants answering each response category (*No* < *Understands* < *Understands and Says*) using a Bayesian, multilevel, ordinal regression model. This model allowed us to estimate both item and participant word-acquisition trajectories, while estimating the effect of our variables of interest: *Age* (number of months elapsed between participants' birth date and questionnaire completion), *Length* (number of phonemes in the X-SAMPA phonological transcription of the word-form), *Exposure* (a language exposure-weighted lexical frequency), and *Cognateness* (defined as the phonological similarity between translation equivalents). We added these variables as main effects, together with the two-way and three-way interactions between *Age*, *Exposure*, and *Cognateness*. Participant-level and item-level random intercepts and slopes were included where appropriate, according to the structure of the data (Barr et al., 2013). We specified a weakly informative prior around the parameters of the model. Equation 1 shows a detailed description of the model.

$$
\begin{aligned}
& \textbf{Response } (k) \textbf{ to word } i \textbf{ by participant } j \\
\text{Response}_{ij} &\sim \text{Cumulative logit}(\theta_{k_{ij}}) \\
& \text{where } k \in \{\text{No} \rightarrow \text{Understands}, \text{Understands} \rightarrow \text{Understands and Says}\} \\
& \textbf{Distribution parameters} \\
\theta_{k_{ij}} &= (\beta_{0_k} + u_{0_{i_k}} + w_{0_{j_k}}) + (\beta_1 + u_{1_i} + w_{1_j}) \cdot \text{Age}_i + \\
& (\beta_2 + u_{2_i} + w_{2_j}) \cdot \text{Length}_{ij} + (\beta_3 + u_{3_i} + w_{3_j}) \cdot \text{Exposure}_{ij} + \\
& (\beta_4 + u_{4_i}) \cdot \text{Cognateness}_{ij} + (\beta_5 + u_{5_i} + w_{3_j}) \cdot (\text{Age}_i \times \text{Exposure}_{ij}) + \\
& (\beta_6 + u_{6_i}) \cdot (\text{Age}_i \times \text{Cognateness}_{ij}) + \\
& (\beta_7 + u_{7_i}) \cdot (\text{Exposure}_{ij} \times \text{Cognateness}_{ij}) \\
& (\beta_8 + u_{8_i}) \cdot (\text{Age}_i \times \text{Exposure}_{ij} \times \text{Cognateness}_{ij}) \\
& \text{where:} \\
& u_{1-8_i} : \text{participant-level adjustments} \\
& w_{1-3_j} : \text{TE-level adjustments} \\
& \textbf{Prior} \\
\beta_{0_k} &\sim \mathcal{N}(-0.25, 0.5);\ \beta_{1-5} \sim \mathcal{N}(0, 1) \\
\sigma_{u_{0-8}, w_{0-3}} &\sim \mathcal{N}_+(1, 0.25);\ \rho_{u_{0-8}, w_{0-3}} \sim \text{LKJcorr}(2) \\
& \text{where } \rho_{u_{0-8}, w_{0-3}} \text{ are the correlations between group-level adjustments}
\end{aligned}
\tag{1}
$$

2.3.3 Model predictors

We developed the *Exposure* predictor to account for the fact that bilinguals' exposure to a given word-form is not only a function of the word-form's lexical frequency, but also of the quantitative input they receive from the language such word-form belongs to, we expressed lexical frequencies as the product between both variables. First, we extracted the child-directed lexical frequency of each word-form from the CHILDES database (MacWhinney, 2000; Sanchez et al., 2019). Using the corresponding lexical frequencies directly from Catalan and Spanish was not possible due to the low number of Catalan participants and tokens available in their corresponding CHILDES corpora, so they were extracted from the English corpora instead. We mapped the lexical frequencies of the English words to their Catalan and Spanish translations (see Fourtassi et al., 2020 for a similar approach), and transformed them to Zipf scores (Van Heuven et al., 2014; Zipf, 1949). We multiplied the resulting lexical frequencies by the reported degree of exposure of the child to Catalan or Spanish. For instance, for a child whose degree of exposure is 80% for Catalan and 20% for Spanish, the expected *Exposure* score to the Catalan word-form *cotxe* [*car*]—with a lexical frequency of 6.33—would be 5.06, while that of its translation to Spanish *coche* would be 1.27.

We defined *Cognateness* as the phonological similarity between each word-form and its translation. For each translation equivalent, we used the `stringdist` (Loo, 2014) R package to calculate the Levenshtein distance between the Catalan and the Spanish phonological transcriptions of the word-forms. The Levenshtein distance measures the number of editions (insertions, deletions, or substitutions) that one string of characters must go through to become identical to the other (Levenshtein, 1966). We divided the Levenshtein distance of each translation equivalent by the length of the longest word-form to correct for word-form length (longer strings are likely to show a larger number of mismatches). Finally, we subtracted the result from one so that it could be interpreted in terms of phonological similarity, instead of phonological distance. This led to a distance metric that ranged from zero to one, where zero indicates that both word-forms are completely different (e.g., /ˈtaw.lɐ/–/ˈme.sa/, *table*), and one indicates that the two word-forms are identical (e.g., /ˈmar/–/ˈmar/, *sea*) [see Floccia et al. (2018); Heeringa & Gooskens (2003); Schepens et al. (2012); Fourtassi et al. (2020); for similar approaches]. Predictors were standardised before entering the model by subtracting the mean of the predictor from each value and dividing the result by the standard deviation of the predictor.

2.3.4 Statistical inference

We assessed the practical relevance of the estimated regression coefficients of the model model following Kruschke & Liddell (2018). First, we specified a region of practical equivalence (ROPE) from -0.025 to +0.025, in the probability scale. This region indicates the range of values that we considered equivalent to zero. We then summarised the posterior distribution of each regression coefficient with the 95% highest density interval (HDI). This interval contains the true value of this coefficient with 95% probability, given the data. Finally, we calculated the proportion of posterior samples in the 95% HDI that fell into the ROPE, noted as $p(\text{ROPE})$, which indicates the probability that the true value of the regression coefficient falls into the ROPE (and therefore should be considered equivalent to zero). For example, $p(\text{ROPE}) = .80$ indicates that, given our data, there is a 80% probability that the true value of the coefficient falls within the ROPE, and can therefore be considered equivalent to zero.

We implemented the model using `brms` (Bürkner, 2017), a R interface to the Stan probabilistic language (2.32.1) (Carpenter et al., 2017). We ran four iteration chains using the by-default No U-Turn Sampler algorithm with 1,000 iterations each and an additional 1,000 warm-up iterations per chain.

3 Results

Chain convergence diagnostics showed adequate values, and the model posterior showed little evidence of correlation, as indicated by negligible pairwise correlation between the marginal posterior distribution of coefficients. Table 3 shows the summary of the posterior distribution of the fixed regression coefficients, and their degree of overlap with the ROPE. For interpretability, we report the estimated regression coefficients transformed to the probability scale[2]. The resulting values correspond to the maximum difference in probability of acquisition (*Comprehension* or *Comprehension and Production*) that corresponds to a one standard deviation change in each predictor.

The coefficient of *Age* showed the strongest association with the probability of acquisition (β = 0.405, 95% HDI = [0.357, 1.806]), with all posterior samples falling out of the ROPE. A one-month increment in age increased a maximum of 0.08 the probability of acquisition. Similarly, the word-form exposure index (*Exposure*) had a strong effect on the probability of acquisition (β = 0.233, 95% HDI = [0.201, 1.074]). All of the posterior samples of this regression coefficient excluded the ROPE. The impact of this predictor on the probability of acquisition was positive: for every standard deviation increase in exposure, the participant was 0.129 more likely to acquire it. Word-form length also showed a significant association with probability of acquisition (β = -0.062, 95% HDI = [-0.086, -0.143]). For every phoneme in the word-form, participants were -0.04 less likely to know it. The 95% HDI of the regression coefficient of the *Age* $\times$ *Exposure* interaction also excluded the ROPE (β = 0.071, 95% HDI = [0.039, 0.414]), showing that the effect of the word-form exposure index differed across ages: older children were more likely to acquire words with a higher exposure rate than younger children.

The posterior distribution of the main effect of cognateness excluded the ROPE completely (β = 0.058, 95% HDI = [0.014, 0.416]). For every 10% increment in cognateness, the acquisition of a word increased in 0.006. The effect of *Cognateness* interacted with that of *Exposure*: the 95% HDI of the regression coefficient of interaction excluded the ROPE entirely (β = -0.06, 95% HDI = [-0.069, -0.184]), suggesting that the effect of cognateness on a word's

[2]The logit and probability scales relate non-linearly. This means that one logit difference is not necessarily translated to a unique value in the probability scale. For example, the probability of acquisition of a given word might increase in 5% when age increases from 22 to 23 months, the probability of acquisition of the same word might only increase in 0.2% when age increases from 31 to 32 months. The linear growth of the probability of acquisition differs along the logistic curve, and therefore deciding the age point at which to report the estimates of the regression coefficients in the probability scale is not trivial. Following Gelman et al. (2020), we report the maximum value of such coefficient, which corresponds to the linear growth (i.e. derivative) of the logistic curve at the age at which most participants were acquiring a given word. This value can be approximated by dividing the coefficient in the logit scale by four: $\hat{\beta}_j/4$, where $\hat{\beta}_j$ is the estimated mean of the posterior distribution of coefficient j.

Table 3: Posterior distribution of regression coefficients. Median: median of the posterior distribution in the probability scale. 95% HDI: 95% highest density interval of the distribution. p(ROPE): overlap between the 95% HDI and the ROPE, indicating the posterior probability that the true value of the coefficient is equivalent to zero.

	β	95% HDI	p(ROPE)
Intercepts (at 22 months)			
Comprehension and Production	0.438	[0.379, 0.496]	-
Comprehension	0.936	[0.92, 0.949]	-
Slopes (upper bound)			
Length (+1 SD, 1.56 phonemes)	-0.062	[-0.086, -0.036]	.000
Age (+1 SD, 4.86 months)	0.405	[0.357, 0.451]	.000
Exposure (+1 SD, 1.81)	0.233	[0.201, 0.268]	.000
Cognateness (+1 SD, 0.26)	0.058	[0.014, 0.104]	.120
Exposure × Cognateness	-0.057	[-0.069, -0.046]	.000
Age × Exposure	0.071	[0.039, 0.104]	.000
Age × Cognateness	0.014	[0, 0.026]	.974
Age × Exposure × Cognateness	-0.018	[-0.027, -0.01]	.907

probability of acquisition changed depending on participants' exposure to the word-form. Follow-up analyses on this interaction showed that when exposure rate was low (e.g., -1 SD), cognateness increased the probability of acquisition substantially. This effect was negligible for words with median or high exposure (+1 SD) (see Figure 1).

An additional analysis including lexical frequency and language exposure as separate predictors (instead of the composite *Exposure* measure) showed equivalent results (see Appendix A). To rule out the possibility that cognateness facilitation effect we found was due to cognateness comprising more frequent syllables than non-cognates—and therefore not because of their cognate status itself—, we compared the syllabic frequency of cognates and non-cognates included in our analyses. To calculate syllable frequency, we first extracted all syllables embedded in the selected words. For each syllable, we summed the lexical frequency of all the words in which such syllable appeared. The resulting value provided an estimate of the number of times the syllable appears in child-directed speech, embedded within different words. Finally, for each word-form, we summed the frequency of its syllables, as an estimate of the syllabic frequency of the word-form. We fit a Bayesian model with *Cognateness* as response variable, and the main effects of syllable frequency and number of syllables (to control for the fact words with more syllables are more likely to score higher in syllabic frequency) as predictors. This model provided strong evidence for the association between cognateness and syllabic frequency being equivalent to zero (see Appendix B).

4 Discussion

This study investigated the impact of cognateness (i.e., phonological similarity between translation equivalents) on the early bilingual lexicon. We used Bayesian Item Response Theory to model the acquisition trajectories of a large sample of Catalan and Spanish words, estimating the effect of cognateness on the probability of acquisition. This model corrected for participants' age, word-form length (number of phonemes), and a novel measure of participants' exposure rate to each word-form. Exposure rates were calculated as a language exposure-weighted lexical frequency score in which each word-form's lexical frequency was corrected by the degree to which the participant was exposed to each language. Overall, we found that cognates (i.e., phonologically similar translation equivalents) were acquired earlier than non-cognates. This effect was mediated by exposure rate: low-exposure word-forms benefited from their cognate status, whereas high-exposure word-forms did not. Using the concept of accumulator (see Kachergis et al., 2022 for review), we provide a theoretical account of bilingual lexical acquisition. In the present account, parallel activation of the two languages plays a central role during the acquisition of early representations in the bilingual lexicon, and in which the dynamics of co-activation between translation equivalents results in an earlier age-of-acquisition.

The present investigation is particularly relevant in the light of two previous findings. First, Floccia et al. (2018) reported that bilingual toddlers learning two typologically close languages (e.g. shared many cognates, like., English-Dutch) showed larger vocabulary sizes than those learning typologically distant languages (e.g. shared fewer cognates, like English-Mandarin). Second, Mitchell et al. (2022) found an earlier age-of-acquisition for cognates, compared to non-cognates. The outcomes of both studies pointed to cognateness facilitating word acquisition through parallel activation. But the underpinnings of such effect were unclear: while parallel activation has been extensively described in experimental studies, current paradigms of bilingual word acquisition and word learning are, to a large extent, dissociated from the mechanisms proposed by previous work on word processing. The notion of *accumulator*,

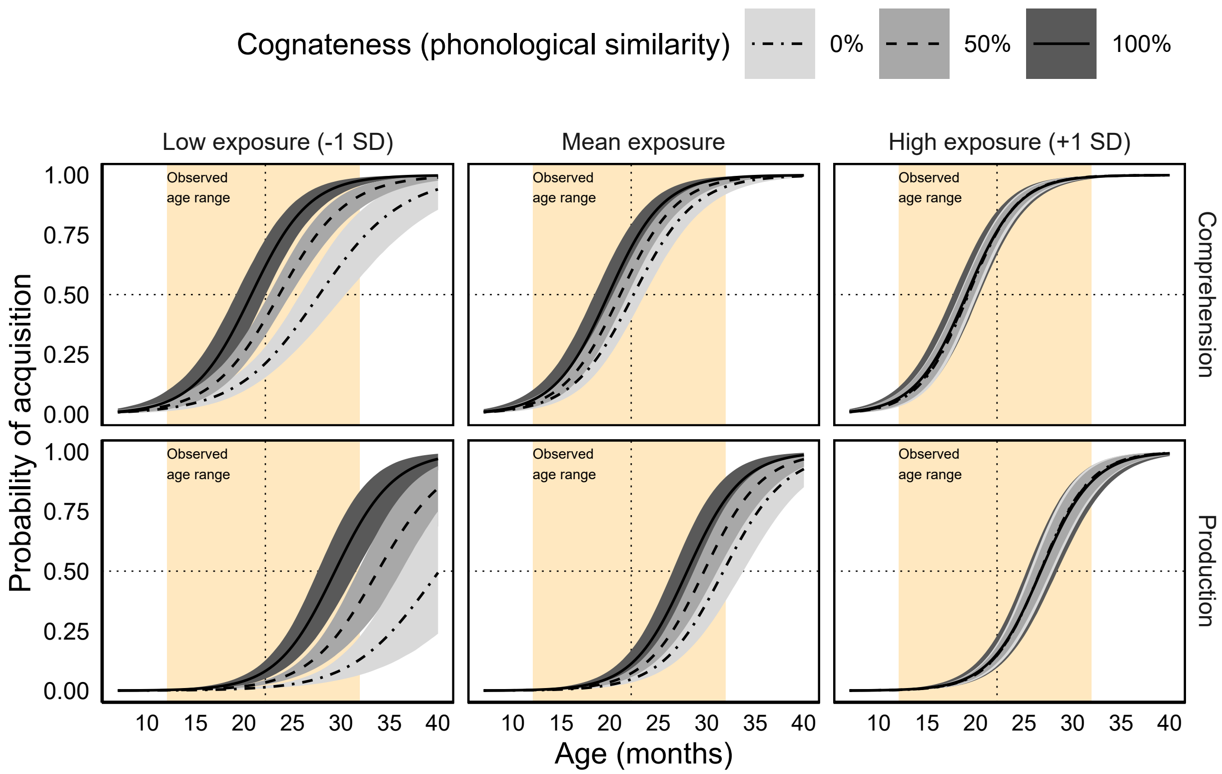

Figure 1: Posterior marginal effects. Lines and error bands correspond to the mean and 95% credible interval of the posterior-predicted means. Different colours indicate different levels of cognateness (phonological similarity). Predictions are presented separately for different degrees of word-form exposure index: little exposure to the word-form, mean exposure, and high exposure). Predictions for Comprehension are shown on top and predictions for Comprehension and Production are shown on the bottom. In-sample predictions lie inside the grey rectangles.

as conceptualised by accumulator models of language acquisition, may provide a convenient theoretical framework to narrow this gap.

Accumulator models devise word acquisition as a continuous process in which the child gathers information about words by accumulating learning instances with such words. When the number of cumulative learning instances for a word reaches some theoretical threshold, the child is considered to have acquired such word. The rate at which a child accumulates learning instances with a word is a function of child-level properties (e.g., ability, amount of quantitative language exposure) and word-level properties (e.g., lexical frequency) (Hidaka, 2013). Through statistical inference, formalised accumulator models provide meaningful information about parameters of interest like the aforementioned predictors (Kachergis et al., 2022; Mollica & Piantadosi, 2017), and allow to generate quantitative predictions about age-of-acquisition and vocabulary growth under competing theoretical accounts (Hidaka, 2013; McMurray, 2007). Using the notion of accumulator, we extended this type of account to the bilingual case. We suggested that the cognate facilitation effect on bilingual word acquisition is the result of cognate words being activated more strongly by their translation than non-cognates. This would lead cognate words to accumulate learning instances at a faster rate than non-cognate words. When a bilingual child is exposed to a word-form, they activate not only its corresponding lexical representation, but also the lexical representation of its translation. The amount of co-activation that spreads from the spoken word-form to its translation is proportional to the amount of phonological similarity between both word-forms. Cognates would receive more activation from their translation than non-cognates, leading children to accumulate learning instances with cognate words at a faster rate than with non-cognate words. As a result, lexical representations of cognate words would consolidate at earlier ages than those of non-cognate words.

These predictions address a critical subject in bilingualism research: do bilingual infants accumulate learning experiences in both languages independently, or does exposure to one language impact the acquisition trajectory of the other language? In the context of lexical acquisition, the former scenario predicts that every learning instance for a given word-form contributes to the acquisition of the representation of such word in the lexicon, while the acquisition of its translation remains unaffected by such experience. In the latter scenario, a learning instance to the same word-form would contribute not only to the acquisition of the representation of the word, but also, to some extent, to the acquisition of its translation. Our findings provide strong support for an account of bilingual vocabulary growth in which the experience and learning outcomes accumulated by the child in one language impact those in the other language through cross-language phonological associations. Such a facilitatory mechanism might be an important piece in the puzzle of bilingual language acquisition. In particular, it may shed some light on why bilingual infants do not show relevant delays in language acquisition milestones compared to their monolingual peers, while receiving a reduced quantity of speech input in each of their languages. Infants in the present study benefited more strongly from the cognateness facilitation effect when acquiring words from the language of lower exposure than in the language of higher exposure.

This mechanism might be extended to provide a plausible explanation for the language similarity facilitation reported by Floccia et al. The authors observed a facilitation in the additional (non-English) language: children learning two typologically close languages knew more words in the additional language than those learning two typologically more distant languages. In their sample, the additional language was consistently also the lower-exposure language for most children, while English was the higher-exposure language. Given that words in English were more likely to be acquired first, higher phonological overlap for words in the language of lower exposure (especially those of lower lexical frequency) would facilitate vocabulary growth for languages sharing more cognates with English.

The asymmetric facilitation of cognateness on word acquisition reported in the present study parallels previous findings in toddlers and adults. For instance, unbalanced (or low-proficiency) bilinguals benefit from cross-language forward priming (dominant to non-dominant) during word processing (De Groot & Nas, 1991; Grainger, 1998; Shook & Marian, 2019; Singh, 2014; Von Holzen et al., 2019; but see Jardak & Byers-Heinlein, 2019). One the other hand, backward priming (non-dominant to dominant) seems less robust and more challenging to detect (e.g., Hoshino et al., 2010; Midgley et al., 2009; but see Duyck & Warlop, 2009). Balanced (or high-proficiency bilinguals) show an equivalent priming facilitation in both directions (Basnight-Brown & Altarriba, 2007; Duñabeitia et al., 2009). These results have been taken as evidence for an asymmetry in the strength of forward and backward connections in the unbalanced bilingual lexicon. Although implemented in different ways, or found under different assumptions, such a dominance-mediated asymmetry is accounted for by multiple models of lexical processing like the Revised Hierarchical Model (Kroll & Stewart, 1994), BIA/BIA+ (Dijkstra & Van Heuven, 2013), BLINCS (Shook & Marian, 2013), or Multilink (Dijkstra et al., 2019), and also by models providing a more development-oriented perspective, like the Ontogenic Model (Bordag et al., 2022; Cook et al., 2016), and BIA-d (Grainger et al., 2010). Overall, this provides an apparently convenient account for the interaction between language dominance and cognateness found in the present study. These models are aimed at explaining results in adults, and their predictions should be taken with caution when extended to early language acquisition.

In adult bilingual populations, language dominance and proficiency are frequently defined using dimensions other than degree of exposure, which is a more common practice in infant research (Marian & Hayakawa, 2021; Rocha-Hidalgo & Barr, 2023). For instance, low-proficiency bilinguals in many of the aforementioned studies acquired their second language years after their toddlerhood. We identify three critical ways in which this prevents a clear comparison between our results and those from studies on second language acquisition in adults. First, in adult second language acquisition, the acquisition of the phonology of the new language must be negotiated with the already acquired phoneme inventory of the first language (e.g., Cutler et al., 2006; Sebastian-Gallés et al., 2006), in place around the first year of life (see Werker & Hensch, 2015 for review). Second, adults acquiring a second language already possess a system of form-meaning mappings, whereas simultaneous bilingual infants must build a lexicon for two languages in the absence of clear form-meaning mappings. Third, adults are assumed to be literate and to possess an orthographic system in place, which may shape how new words are integrated in the lexicon and processed during experimental tasks (e.g., Thierry & Wu, 2007). In this scenario, the acquisition of a second language may take place in a substantially different way compared to how bilingual infants acquire two languages from birth. A more similar case to the one concerning the present study is considered by the DevLex-II model (Zhao & Li, 2010), which captures unique features of the early bilingual lexicon, and considers the case of infants simultaneously acquiring their two language. In line with the adult models, DevLex-II predicts asymmetries between word representation from the dominant and the non-dominant language. Simulations from DevLex-II result in an asymmetric cross-language priming, in which words from the dominant (acquired acquired) language primed more strongly the recognition of words in the non-dominant language (later acquired) than in the other direction (Zhao & Li, 2013).

In summary, there is a compelling case for attributing asymmetric effects of parallel activation to differences in activation strength between forward and backward connections. It is nonetheless possible that, as argued in the introduction, the asymmetric effect of cognateness found in the present study is simply the result of infants being exposed more frequently to words in the dominant language than to words in the non-dominant language. This would lead to words in the non-dominant language receiving additional parallel activation, compared to words in the dominant language, and therefore benefiting more strongly from their cognate status. These two accounts are not mutually exclusive, as words in the dominant language may active more *strongly* their translations than *vice versa*, on top of such activation being more frequent. Further research is needed in order to clarify this issue.

It might be argued that our results reflect the fact that cognate translation equivalents are represented in the initial bilingual lexicon as the *same* lexical entry. Because cognates correspond to similar sounding word-forms in equivalent referential contexts (e.g., hearing /ˈgat/ and /ˈga.to/ in the same situations), it is possible that infants classify both are as acceptable variations of the same word-form, therefore treating them as a single lexical item. This would lead to a faster increase in cumulative learning instances, and to earlier ages of acquisition for cognate translation equivalents (for which listening to each word-form contributes to the acquisition of its shared representation), compared to non-cognates (for which listening to each word-form contributes to the acquisition of a separate representation). This mechanism could potentially explain the earlier age-of-acquisition effect of cognates found in the present study, without the need of parallel activation playing any relevant role. Mitchell et al. (2022) discuss this possibility as a possible explanation of the cognate facilitation effect, in which bilinguals only need to map one word-form to the referent in the case of cognates, while mapping two distinct word-forms in the case of non-cognates. However, previous work on mispronunciation perception and learning of minimal pair words points in a different direction. Bilingual toddlers show monolingual-level sensitivity to slight phonetic changes in a word-form, according to their performance in word recognition tasks (Bailey & Plunkett, 2002; Mani & Plunkett, 2011; Ramon-Casas et al., 2009, 2017; Ramon-Casas & Bosch, 2010; Swingley, 2005; Swingley & Aslin, 2000; Tamási et al., 2017; Wewalaarachchi et al., 2017). The ability to differentiate between similar-sounding word-forms is also reflected in word learning, as bilinguals seem to be able to map minimal pairs to distinct referents (Havy et al., 2016; Mattock et al., 2010; Ramon-Casas et al., 2017). Overall, it seems that bilinguals consider small differences in the phonological forms of words as relevant at the lexical level. We argue that this shows evidence that bilingual toddlers likely form distinct lexical representations for even near-identical cognates.

Our study shares similar methodological limitations with previous work using vocabulary reports provided by caregivers. Such reports can be subject to measurement error induced by caregivers who may sometimes overestimate or underestimate participants' true probability of word acquisition (e.g., Houston-Price et al., 2007). In the case of bilingual research additional biases may be in place. Although in the present study caregivers were explicitly instructed *not* to rely on their responses to Catalan words when responding to Spanish (and vice versa), it is possible that some caregivers assumed—at least to some extent—that because the child knew a word in one language, the child should also know the word in the other language. This bias would especially affect similar-sounding words, i.e., cognates. Production estimates may be more prompt to such biases, in part because of the slower pace at which infants' articulatory abilities develop, compared to their word recognition abilities (Hustad et al., 2021). This gap between comprehension and production is even larger in the less dominant language of bilingual children (Giguere & Hoff, 2022). For this reason, caregivers may be more uncertain about what words can be counted as *acquired* in this modality. Despite such potential biases, vocabulary checklist filled by parents show strong evidence of concurrent validity with other estimates of vocabulary size or lexical processing (Feldman et al., 2005; Gillen et al., 2021; but see Houston-Price et al., 2007).

The present study contributes with a specific data point to the complex landscape of bilingualism research. Bilinguals are a remarkably heterogeneous population difficult to be satisfactorily characterised in a comprehensive way (Sebastian-Galles & Santolin, 2020). Bilinguals differ across multiple dimensions. Such differences span from exclusively linguistic factors; such as the amount of overlap between the phonemic inventories of the two languages being learned (e.g., low, like the case of English and Mandarin, or high like the case of Spanish and Greek), to extralinguistic factors like the socio-linguistic situation in which the two languages co-exist (e.g., in some regions both languages are co-official and used in similar contexts, while in others, one of the languages hardly has any societal presence, i.e., heritage languages). This diversity of situations in which bilingual toddlers acquire language calls for special consideration of the generalisability of results in bilingualism research. Our sample, although homogeneous (e.g., similar parental educational level across), represents a particular bilingual sociolinguistic environment: the languages involved in the present investigation, Catalan and Spanish, co-exist in Catalonia as official languages, both languages are used in fairly similar contexts, and both languages are known by the majority of the population. In 2018, more than 81.2% of a representative sample of 8,780 adults aged 15 years or older living in Catalonia reported being able to speak Catalan, and more than 99.5% of the same population reported being able to speak Spanish (*Els Usos Lingüístics de La Població de Catalunya*, 2018). In addition, Catalan and Spanish are Romance languages and share a considerable amount of cognates. Extending our analyses to other bilingual populations learning typologically more distant languages, and whose languages tend to be used in more distinct contexts (e.g., heritage languages) should be a natural future step for the present investigation.

To conclude, our study provides novel insights about word acquisition in bilingual contexts, and how the presence of cognates in the children's linguistic input impacts the early formation of the lexicon. We found that during the acquisition of low frequency words, bilingual children seem to benefit more strongly from the word-forms's phonological similarity with its translation in the other language. Capitalising on the notion of *accumulator* of linguistic input, we put forward a theoretical account of bilingual word learning, in which cognateness interacts with lexical frequency and language exposure to boost the acquisition of translation equivalents.

References

Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). Formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using r. *Behavior Research Methods*, *52*(1), 376–387. https://doi.org/10.3758/s13428-019-01236-y

Bailey, T. M., & Plunkett, K. (2002). Phonological specificity in early words. *Cognitive Development*, *17*(2), 1265–1282. https://doi.org/10.1016/S0885-2014(02)00116-8

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Basnight-Brown, D. M., & Altarriba, J. (2007). Differences in semantic and translation priming across languages: The role of language direction and language dominance. *Memory & Cognition*, *35*, 953–965.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Bergelson, E., & Swingley, D. (2015). Early word comprehension in infants: Replication and extension. *Language Learning and Development*, *11*(4), 369–380. https://doi.org/10.1080/15475441.2014.979387

Blom, E., Boerma, T., Bosma, E., Cornips, L., Heuij, K. van den, & Timmermeister, M. (2020). Cross-language distance influences receptive vocabulary outcomes of bilingual children. *First Language*, *40*(2), 151–171. https://doi.org/10.1177/0142723719892794

Bordag, D., Gor, K., & Opitz, A. (2022). Ontogenesis model of the L2 lexical representation. *Bilingualism: Language and Cognition*, *25*(2), 185–201.

Bosch, L., & Ramon-Casas, M. (2014). First translation equivalents in bilingual toddlers' expressive vocabulary: Does form similarity matter? *International Journal of Behavioral Development*, *38*(4), 317–322. https://doi.org/10.1177/0165025414532559

Bosma, E., Blom, E., Hoekstra, E., & Versloot, A. (2019). A longitudinal study on the gradual cognate facilitation effect in bilingual children's frisian receptive vocabulary. *International Journal of Bilingual Education and Bilingualism*, *22*(4), 371–385. https://doi.org/10.1080/13670050.2016.1254152

Bosma, E., & Nota, N. (2020). Cognate facilitation in frisian–dutch bilingual children's sentence reading: An eye-tracking study. *Journal of Experimental Child Psychology*, *189*, 104699.

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan : A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Cook, S. V., Panda, N. B., Lancaster, A. K., & Gor, K. (2016). Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology*, *7*, 1345.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1283–1296. https://doi.org/10.1037/0278-7393.26.5.1283

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, *34*(2), 269–284.

De Groot, A. M., & Nas, G. L. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *Journal of Memory and Language*, *30*(1), 90–123.

Dijkstra, T., & Van Heuven, W. J. (2013). The BIA model and bilingual word recognition. In *Localist connectionist approaches to human cognition* (pp. 189–225). Psychology Press.

Dijkstra, T., Wahl, A., Buytenhuijs, F., Halem, N. V., Al-Jibouri, Z., Korte, M. D., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, *22*(4), 657–679. https://doi.org/10.1017/S1366728918000287

Duñabeitia, J. A., Perea, M., & Carreiras, M. (2009). Masked translation priming effects with highly proficient simultaneous bilinguals. *Experimental Psychology*.

Duyck, W., & Warlop, N. (2009). Translation priming between the native language and a second language: New evidence from dutch-french bilinguals. *Experimental Psychology*, *56*(3), 173–179.

Els usos lingüístics de la població de catalunya. (2018). Generalitat de Catalunya. https://llengua.gencat.cat/web/.content/documents/dadesestudis/altres/arxius/dossier-eulp-2018.pdf

Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs-Lasky, M., Rockette, H. E., & Paradise, J. L. (2005). Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. *Child Development*, *76*(4), 856–868. https://doi.org/10.1111/j.1467-8624.2005.00882.x

Fenson, L. et al. (2007). *MacArthur-bates communicative development inventories*. Paul H. Brookes Publishing Company Baltimore, MD.

Floccia, C., Delle Luche, C., Lepadatu, I., Chow, J., Ratnage, P., & Plunkett, K. (2020). Translation equivalent and cross-language semantic priming in bilingual toddlers. *Journal of Memory and Language*, *112*, 104086.

Floccia, C., Sambrook, T. D., Delle Luche, C., Kwok, R., Goslin, J., White, L., Cattani, A., Sullivan, E., AbbotSmith, K., Krott, A., Mills, D., Rowland, C., Gervain, J., & Plunkett, K. (2018). I: introduction. *Monographs of the Society for Research in Child Development*, *83*(1), 7–29. https://doi.org/10.1111/mono.12348

Fourtassi, A., Bian, Y., & Frank, M. C. (2020). The growth of children's semantic and phonological networks: Insight from 10 languages. *Cognitive Science*, *44*(7), e12847. https://doi.org/10.1111/cogs.12847

Gampe, A., Quick, A. E., & Daum, M. M. (2021). Does linguistic similarity affect early simultaneous bilingual language acquisition? *Journal of Language Contact*, *13*(3), 482–500.

Garcia-Castro, G., Ávila-Varela, D. S., & Sebastian-Galles, N. (2023). *Bvq: Barcelona vocabulary questionnaire database and helper functions*. https://gongcastro.github.io/bvq

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

Giguere, D., & Hoff, E. (2022). Bilingual development in the receptive and expressive domains: They differ. *International Journal of Bilingual Education and Bilingualism*, *25*(10), 3849–3858. https://doi.org/10.1080/13670050.2022.2087039

Gillen, N. A., Siow, S., Lepadatu, I., Sucevic, J., Plunkett, K., & Duta, M. (2021). *Tapping into the potential of remote developmental research: Introducing the OxfordBabylab app*. PsyArXiv. https://doi.org/10.31234/osf.io/kxhmw

Gimeno-Martínez, M., Mädebach, A., & Baus, C. (2021). Cross-linguistic interactions across modalities: Effects of the oral language on sign production. *Bilingualism: Language and Cognition*, *24*(4), 779–790. https://doi.org/10.1017/S1366728921000171

Gonzalez-Barrero, A. M., Schott, E., & Byers-Heinlein, K. (2020). *Bilingual adjusted vocabulary: A developmentally-informed bilingual vocabulary measure*. PsyArXiv. https://doi.org/10.31234/osf.io/x7s4u

Grainger, J. (1998). Masked priming by translation equivalents in proficient bilinguals. *Language and Cognitive Processes*, *13*(6), 601–623.

Grainger, J., Midgley, K., & Holcomb, P. J. (2010). Re-thinking the bilingual interactive-activation model from a developmental perspective (BIA-d). *Language Acquisition Across Linguistic and Cognitive Systems*, *52*, 267–283.

Grosjean, F. (2021). The extent of bilingualism. *Life as a Bilingual*, 27–39.

Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a british communicative development inventory. *Journal of Child Language*, *27*(3), 689–705.

Havy, M., Bouchon, C., & Nazzi, T. (2016). Phonetic processing when learning words: The case of bilingual infants. *International Journal of Behavioral Development*, *40*(1), 41–52. https://doi.org/10.1177/0165025415570646

Heeringa, W., & Gooskens, C. (2003). Norwegian dialects examined perceptually and acoustically. *Computers and the Humanities*, *37*(3), 293–315. https://doi.org/10.1023/A:1025087115665

Hidaka, S. (2013). A computational model associating learning process, word attributes, and age of acquisition. *PLOS ONE*, *8*(11), e76242. https://doi.org/10.1371/journal.pone.0076242

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development*. *Journal of Child Language*, *39*(1), 1–27. https://doi.org/10.1017/S0305000910000759

Hoshino, N., & Kroll, J. F. (2008). Cognate effects in picture naming: Does cross-language activation survive a change of script? *Cognition*, *106*(1), 501–511. https://doi.org/10.1016/j.cognition.2007.02.001

Hoshino, N., Midgley, K. J., Holcomb, P. J., & Grainger, J. (2010). An ERP investigation of masked cross-script translation priming. *Brain Research*, *1344*, 159–172.

Houston-Price, C., Mather, E., & Sakkalou, E. (2007). Discrepancy between parental reports of infants' receptive vocabulary and infants' behaviour in a preferential looking task. *Journal of Child Language*, *34*(4), 701–724. https://doi.org/10.1017/S0305000907008124

Hustad, K. C., Mahr, T. J., Natzke, P., & Rathouz, P. J. (2021). Speech development between 30 and 119 months in typical children i: Intelligibility growth curves for single-word and multiword productions. *Journal of Speech, Language, and Hearing Research*, *64*(10), 3707–3719. https://doi.org/10.1044/2021_JSLHR-21-00142

Jardak, A., & Byers-Heinlein, K. (2019). Labels or concepts? The development of semantic networks in bilingual two-year-olds. *Child Development*, *90*(2), e212–e229.

Kachergis, G., Marchman, V. A., & Frank, M. C. (2022). Toward a "standard model" of early language learning. *Current Directions in Psychological Science*, *31*(1), 20–27. https://doi.org/10.1177/09637214211057836

Kroll, J. F., & Ma, F. (2017). The bilingual lexicon. *The Handbook of Psycholinguistics*, 294–319.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connection between bilingual memory representations. *Journal of Memory and Language*, *33*(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

Kruschke, J. K., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and planning from a bayesian perspective. *Psychonomic Bulletin &Review*, *25*, 178–206. https://doi.org/10.3758/s13423-016-1221-4

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady*, *10*, 707–710.

Loo, M. P. J. van der. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*(1), 111–122. https://doi.org/10.32614/RJ-2014-011

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

Mani, N., & Plunkett, K. (2011). Does size matter? Subsegmental cues to vowel mispronunciation detection*. *Journal of Child Language*, *38*(3), 606–627. https://doi.org/10.1017/S0305000910000243

Marian, V., & Hayakawa, S. (2021). Measuring bilingualism: The quest for a "bilingualism quotient." *Applied Psycholinguistics*, *42*(2), 527–548.

Mattock, K., Polka, L., Rvachew, S., & Krehm, M. (2010). The first steps in word learning are easier when the shoes fit: Comparing monolingual and bilingual infants. *Developmental Science*, *13*(1), 229–243. https://doi.org/10.1111/j.1467-7687.2009.00891.x

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631–631.

Midgley, K. J., Holcomb, P. J., & Grainger, J. (2009). Language effects in second language learners and proficient bilinguals investigated with event-related potentials. *Journal of Neurolinguistics*, *22*(3), 281–300.

Mitchell, L., Tsui, R. K. Y., & Byers-Heinlein, K. (2022). *Cognates are advantaged in early bilingual expressive vocabulary development*. PsyArXiv. https://doi.org/10.31234/osf.io/daktp

Mollica, F., & Piantadosi, S. T. (2017). How data drive early word learning: A cross-linguistic waiting time analysis. *Open Mind*, *1*(2), 67–77. https://doi.org/10.1162/OPMI_a_00006

Morford, J. P., Wilkinson, E., Villwock, A., Piñar, P., & Kroll, J. F. (2011). When deaf signers read english: Do written words activate their sign translations? *Cognition*, *118*(2), 286–292. https://doi.org/10.1016/j.cognition.2010.11.006

Oller, D. K., & Eilers, R. E. (2002). *Language and literacy in bilingual children*. Multilingual Matters.

Patterson, J. L. (2004). Comparing bilingual and monolingual toddlers' expressive vocabulary size. *Journal of Speech, Language, and Hearing Research*, *47*(5), 1213–1215. https://doi.org/10.1044/1092-4388(2004/089)

Patterson, J. L., & Pearson, B. Z. (2004). Bilingual lexical development: Influences, contexts, and processes. In *Bilingual language development and disorders in spanish-english speakers* (pp. 77–104). Paul H. Brookes Publishing Co.

Pearson, B. Z., & Fernández, S. C. (1994). Patterns of interaction in the lexical growth in two languages of bilingual infants and toddlers. *Language Learning*, *44*(4), 617–653. https://doi.org/10.1111/j.1467-1770.1994.tb00633.x

Poarch, G. J., & Hell, J. G. van. (2012). Cross-language activation in children's speech production: Evidence from second language learners, bilinguals, and trilinguals. *Journal of Experimental Child Psychology*, *111*(3), 419–438. https://doi.org/10.1016/j.jecp.2011.09.008

Poulin-Dubois, D., Bialystok, E., Blaye, A., Polonia, A., & Yott, J. (2013). Lexical access and vocabulary development in very young bilinguals. *International Journal of Bilingualism*, *17*(1), 57–70.

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Ramon-Casas, M., & Bosch, L. (2010). Are non-cognate words phonologically better specified than cognates in the early lexicon of bilingual children. *Selected Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology*, 31–36.

Ramon-Casas, M., Fennell, C. T., & Bosch, L. (2017). Minimal-pair word learning by bilingual toddlers: The catalan /e/-// contrast revisited. *Bilingualism: Language and Cognition*, *20*(3), 649–656. https://doi.org/10.1017/S1366728916001115

Ramon-Casas, M., Swingley, D., Sebastián-Gallés, N., & Bosch, L. (2009). Vowel categorization during word recognition in bilingual toddlers. *Cognitive Psychology*, *59*(1), 96–121. https://doi.org/10.1016/j.cogpsych.2009.02.002

Rocha-Hidalgo, J., & Barr, R. (2023). Defining bilingualism in infancy and toddlerhood: A scoping review. *International Journal of Bilingualism*, *27*(3), 253–274.

Samuelson, L. K. (2021). Toward a precision science of word learning: Understanding individual vocabulary pathways. *Child Development Perspectives*, *15*(2), 117–124. https://doi.org/10.1111/cdep.12408

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, *51*(4), 1928–1941.

Schelletter, C. (2002). The effect of form similarity on bilingual children's lexical development. *Bilingualism: Language and Cognition*, *5*(2), 93–107.

Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in europe as based on levenshtein distance. *Bilingualism: Language and Cognition*, *15*(1), 157–166.

Sebastian-Galles, N., & Santolin, C. (2020). Bilingual acquisition: The early steps. *Annual Review of Developmental Psychology*, *2*(1), 47–68. https://doi.org/10.1146/annurev-devpsych-013119-023724

Sebastian-Gallés, N., Rodrguez-Fornells, A., Diego-Balaguer, R. de, & Daz, B. (2006). First-and second-language phonological representations in the mental lexicon. *Journal of Cognitive Neuroscience*, *18*(8), 1277–1291.

Shook, A., & Marian, V. (2012). Bimodal bilinguals co-activate both languages during spoken comprehension. *Cognition*, *124*(3), 314–324.

Shook, A., & Marian, V. (2013). The bilingual language interaction network for comprehension of speech*. *Bilingualism: Language and Cognition*, *16*(2), 304–324. https://doi.org/10.1017/S1366728912000466

Shook, A., & Marian, V. (2019). Covert co-activation of bilinguals' non-target language: Phonological competition from translations. *Linguistic Approaches to Bilingualism*, *9*(2), 228–252.

Singh, L. (2014). One world, two languages: Cross-language semantic priming in bilingual toddlers. *Child Development*, *85*(2), 755–766.

Smithson, L., Paradis, J., & Nicoladis, E. (2014). Bilingualism and receptive vocabulary achievement: Could sociocultural context make a difference? *Bilingualism: Language and Cognition*, *17*(4), 810–821. https://doi.org/10.1017/S1366728913000813

Spivey, M. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science*, *10*(3), 281–284.

Swingley, D. (2005). 11-month-olds' knowledge of how familiar words sound. *Developmental Science*, *8*(5), 432–443. https://doi.org/10.1111/j.1467-7687.2005.00432.x

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–166. https://doi.org/10.1016/S0010-0277(00)00081-0

Tamási, K., McKean, C., Gafos, A., Fritzsche, T., & Höhle, B. (2017). Pupillometry registers toddlers' sensitivity to degrees of mispronunciation. *Journal of Experimental Child Psychology*, *153*, 140–148. https://doi.org/10.1016/j.jecp.2016.07.014

Thierry, G., & Wu, Y. J. (2007). Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences*, *104*(30), 12530–12535. https://doi.org/10.1073/pnas.0609927104

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*(2), 172–175. https://doi.org/10.1111/1467-9280.00127

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for british english. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Von Holzen, K., Fennell, C. T., & Mani, N. (2019). The impact of cross-language phonological overlap on bilingual and monolingual toddlers' word recognition. *Bilingualism: Language and Cognition*, *22*(3), 476–499. https://doi.org/10.1017/S1366728918000597

Von Holzen, K., & Mani, N. (2012). Language nonselective lexical access in bilingual toddlers. *Journal of Experimental Child Psychology*, *113*(4), 569–586. https://doi.org/10.1016/j.jecp.2012.08.001

Wells, J. C. (1995). *Computer-coding the IPA: A proposed extension of SAMPA*. *4*(28), 1995.

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: New directions. *Annual Review of Psychology*, *66*, 173–196.

Wewalaarachchi, T. D., Wong, L. H., & Singh, L. (2017). Vowels, consonants, and lexical tones: Sensitivity to phonological variation in monolingual mandarin and bilingual english–mandarin toddlers. *Journal of Experimental Child Psychology*, *159*, 16–33. https://doi.org/10.1016/j.jecp.2017.01.009

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Zhao, X., & Li, P. (2010). Bilingual lexical interactions in an unsupervised neural network model. *International Journal of Bilingual Education and Bilingualism*, *13*(5), 505–524.

Zhao, X., & Li, P. (2013). Simulating cross-language priming with a dynamic computational model of the lexicon. *Bilingualism: Language and Cognition*, *16*(2), 288–303.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.

Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). Formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using r. *Behavior Research Methods*, *52*(1), 376–387. https://doi.org/10.3758/s13428-019-01236-y

Bailey, T. M., & Plunkett, K. (2002). Phonological specificity in early words. *Cognitive Development*, *17*(2), 1265–1282. https://doi.org/10.1016/S0885-2014(02)00116-8

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Basnight-Brown, D. M., & Altarriba, J. (2007). Differences in semantic and translation priming across languages: The role of language direction and language dominance. *Memory & Cognition*, *35*, 953–965.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Bergelson, E., & Swingley, D. (2015). Early word comprehension in infants: Replication and extension. *Language Learning and Development*, *11*(4), 369–380. https://doi.org/10.1080/15475441.2014.979387

Blom, E., Boerma, T., Bosma, E., Cornips, L., Heuij, K. van den, & Timmermeister, M. (2020). Cross-language distance influences receptive vocabulary outcomes of bilingual children. *First Language*, *40*(2), 151–171. https://doi.org/10.1177/0142723719892794

Bordag, D., Gor, K., & Opitz, A. (2022). Ontogenesis model of the L2 lexical representation. *Bilingualism: Language and Cognition*, *25*(2), 185–201.

Bosch, L., & Ramon-Casas, M. (2014). First translation equivalents in bilingual toddlers' expressive vocabulary: Does form similarity matter? *International Journal of Behavioral Development*, *38*(4), 317–322. https://doi.org/10.1177/0165025414532559

Bosma, E., Blom, E., Hoekstra, E., & Versloot, A. (2019). A longitudinal study on the gradual cognate facilitation effect in bilingual children's frisian receptive vocabulary. *International Journal of Bilingual Education and Bilingualism*, *22*(4), 371–385. https://doi.org/10.1080/13670050.2016.1254152

Bosma, E., & Nota, N. (2020). Cognate facilitation in frisian–dutch bilingual children's sentence reading: An eye-tracking study. *Journal of Experimental Child Psychology*, *189*, 104699.

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan : A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Cook, S. V., Panda, N. B., Lancaster, A. K., & Gor, K. (2016). Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology*, *7*, 1345.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1283–1296. https://doi.org/10.1037/0278-7393.26.5.1283

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, *34*(2), 269–284.

De Groot, A. M., & Nas, G. L. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *Journal of Memory and Language*, *30*(1), 90–123.

Dijkstra, T., & Van Heuven, W. J. (2013). The BIA model and bilingual word recognition. In *Localist connectionist approaches to human cognition* (pp. 189–225). Psychology Press.

Dijkstra, T., Wahl, A., Buytenhuijs, F., Halem, N. V., Al-Jibouri, Z., Korte, M. D., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, *22*(4), 657–679. https://doi.org/10.1017/S1366728918000287

Duñabeitia, J. A., Perea, M., & Carreiras, M. (2009). Masked translation priming effects with highly proficient simultaneous bilinguals. *Experimental Psychology*.

Duyck, W., & Warlop, N. (2009). Translation priming between the native language and a second language: New evidence from dutch-french bilinguals. *Experimental Psychology*, *56*(3), 173–179.

Els usos lingüístics de la població de catalunya. (2018). Generalitat de Catalunya. https://llengua.gencat.cat/web/.content/documents/dadesestudis/altres/arxius/dossier-eulp-2018.pdf

Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs-Lasky, M., Rockette, H. E., & Paradise, J. L. (2005). Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. *Child Development*, *76*(4), 856–868. https://doi.org/10.1111/j.1467-8624.2005.00882.x

Fenson, L. et al. (2007). *MacArthur-bates communicative development inventories*. Paul H. Brookes Publishing Company Baltimore, MD.

Floccia, C., Delle Luche, C., Lepadatu, I., Chow, J., Ratnage, P., & Plunkett, K. (2020). Translation equivalent and cross-language semantic priming in bilingual toddlers. *Journal of Memory and Language*, *112*, 104086.

Floccia, C., Sambrook, T. D., Delle Luche, C., Kwok, R., Goslin, J., White, L., Cattani, A., Sullivan, E., AbbotSmith, K., Krott, A., Mills, D., Rowland, C., Gervain, J., & Plunkett, K. (2018). I: introduction. *Monographs of the Society for Research in Child Development*, *83*(1), 7–29. https://doi.org/10.1111/mono.12348

Fourtassi, A., Bian, Y., & Frank, M. C. (2020). The growth of children's semantic and phonological networks: Insight from 10 languages. *Cognitive Science*, *44*(7), e12847. https://doi.org/10.1111/cogs.12847

Gampe, A., Quick, A. E., & Daum, M. M. (2021). Does linguistic similarity affect early simultaneous bilingual language acquisition? *Journal of Language Contact*, *13*(3), 482–500.

Garcia-Castro, G., Ávila-Varela, D. S., & Sebastian-Galles, N. (2023). *Bvq: Barcelona vocabulary questionnaire database and helper functions*. https://gongcastro.github.io/bvq

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

Giguere, D., & Hoff, E. (2022). Bilingual development in the receptive and expressive domains: They differ. *International Journal of Bilingual Education and Bilingualism*, *25*(10), 3849–3858. https://doi.org/10.1080/13670050.2022.2087039

Gillen, N. A., Siow, S., Lepadatu, I., Sucevic, J., Plunkett, K., & Duta, M. (2021). *Tapping into the potential of remote developmental research: Introducing the OxfordBabylab app*. PsyArXiv. https://doi.org/10.31234/osf.io/kxhmw

Gimeno-Martínez, M., Mädebach, A., & Baus, C. (2021). Cross-linguistic interactions across modalities: Effects of the oral language on sign production. *Bilingualism: Language and Cognition*, *24*(4), 779–790. https://doi.org/10.1017/S1366728921000171

Gonzalez-Barrero, A. M., Schott, E., & Byers-Heinlein, K. (2020). *Bilingual adjusted vocabulary: A developmentally-informed bilingual vocabulary measure*. PsyArXiv. https://doi.org/10.31234/osf.io/x7s4u

Grainger, J. (1998). Masked priming by translation equivalents in proficient bilinguals. *Language and Cognitive Processes*, *13*(6), 601–623.

Grainger, J., Midgley, K., & Holcomb, P. J. (2010). Re-thinking the bilingual interactive-activation model from a developmental perspective (BIA-d). *Language Acquisition Across Linguistic and Cognitive Systems*, *52*, 267–283.

Grosjean, F. (2021). The extent of bilingualism. *Life as a Bilingual*, 27–39.

Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a british communicative development inventory. *Journal of Child Language*, *27*(3), 689–705.

Havy, M., Bouchon, C., & Nazzi, T. (2016). Phonetic processing when learning words: The case of bilingual infants. *International Journal of Behavioral Development*, *40*(1), 41–52. https://doi.org/10.1177/0165025415570646

Heeringa, W., & Gooskens, C. (2003). Norwegian dialects examined perceptually and acoustically. *Computers and the Humanities*, *37*(3), 293–315. https://doi.org/10.1023/A:1025087115665

Hidaka, S. (2013). A computational model associating learning process, word attributes, and age of acquisition. *PLOS ONE*, *8*(11), e76242. https://doi.org/10.1371/journal.pone.0076242

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development*. *Journal of Child Language*, *39*(1), 1–27. https://doi.org/10.1017/S0305000910000759

Hoshino, N., & Kroll, J. F. (2008). Cognate effects in picture naming: Does cross-language activation survive a change of script? *Cognition*, *106*(1), 501–511. https://doi.org/10.1016/j.cognition.2007.02.001

Hoshino, N., Midgley, K. J., Holcomb, P. J., & Grainger, J. (2010). An ERP investigation of masked cross-script translation priming. *Brain Research*, *1344*, 159–172.

Houston-Price, C., Mather, E., & Sakkalou, E. (2007). Discrepancy between parental reports of infants' receptive vocabulary and infants' behaviour in a preferential looking task. *Journal of Child Language*, *34*(4), 701–724. https://doi.org/10.1017/S0305000907008124

Hustad, K. C., Mahr, T. J., Natzke, P., & Rathouz, P. J. (2021). Speech development between 30 and 119 months in typical children i: Intelligibility growth curves for single-word and multiword productions. *Journal of Speech, Language, and Hearing Research*, *64*(10), 3707–3719. https://doi.org/10.1044/2021_JSLHR-21-00142

Jardak, A., & Byers-Heinlein, K. (2019). Labels or concepts? The development of semantic networks in bilingual two-year-olds. *Child Development*, *90*(2), e212–e229.

Kachergis, G., Marchman, V. A., & Frank, M. C. (2022). Toward a "standard model" of early language learning. *Current Directions in Psychological Science*, *31*(1), 20–27. https://doi.org/10.1177/09637214211057836

Kroll, J. F., & Ma, F. (2017). The bilingual lexicon. *The Handbook of Psycholinguistics*, 294–319.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connection between bilingual memory representations. *Journal of Memory and Language*, *33*(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

Kruschke, J. K., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and planning from a bayesian perspective. *Psychonomic Bulletin &Review*, *25*, 178–206. https://doi.org/10.3758/s13423-016-1221-4

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady*, *10*, 707–710.

Loo, M. P. J. van der. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*(1), 111–122. https://doi.org/10.32614/RJ-2014-011

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

Mani, N., & Plunkett, K. (2011). Does size matter? Subsegmental cues to vowel mispronunciation detection*. *Journal of Child Language*, *38*(3), 606–627. https://doi.org/10.1017/S0305000910000243
Marian, V., & Hayakawa, S. (2021). Measuring bilingualism: The quest for a "bilingualism quotient." *Applied Psycholinguistics*, *42*(2), 527–548.
Mattock, K., Polka, L., Rvachew, S., & Krehm, M. (2010). The first steps in word learning are easier when the shoes fit: Comparing monolingual and bilingual infants. *Developmental Science*, *13*(1), 229–243. https://doi.org/10.1111/j.1467-7687.2009.00891.x
McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631–631.
Midgley, K. J., Holcomb, P. J., & Grainger, J. (2009). Language effects in second language learners and proficient bilinguals investigated with event-related potentials. *Journal of Neurolinguistics*, *22*(3), 281–300.
Mitchell, L., Tsui, R. K. Y., & Byers-Heinlein, K. (2022). *Cognates are advantaged in early bilingual expressive vocabulary development*. PsyArXiv. https://doi.org/10.31234/osf.io/daktp
Mollica, F., & Piantadosi, S. T. (2017). How data drive early word learning: A cross-linguistic waiting time analysis. *Open Mind*, *1*(2), 67–77. https://doi.org/10.1162/OPMI_a_00006
Morford, J. P., Wilkinson, E., Villwock, A., Piñar, P., & Kroll, J. F. (2011). When deaf signers read english: Do written words activate their sign translations? *Cognition*, *118*(2), 286–292. https://doi.org/10.1016/j.cognition.2010.11.006
Oller, D. K., & Eilers, R. E. (2002). *Language and literacy in bilingual children*. Multilingual Matters.
Patterson, J. L. (2004). Comparing bilingual and monolingual toddlers' expressive vocabulary size. *Journal of Speech, Language, and Hearing Research*, *47*(5), 1213–1215. https://doi.org/10.1044/1092-4388(2004/089)
Patterson, J. L., & Pearson, B. Z. (2004). Bilingual lexical development: Influences, contexts, and processes. In *Bilingual language development and disorders in spanish-english speakers* (pp. 77–104). Paul H. Brookes Publishing Co.
Pearson, B. Z., & Fernández, S. C. (1994). Patterns of interaction in the lexical growth in two languages of bilingual infants and toddlers. *Language Learning*, *44*(4), 617–653. https://doi.org/10.1111/j.1467-1770.1994.tb00633.x
Poarch, G. J., & Hell, J. G. van. (2012). Cross-language activation in children's speech production: Evidence from second language learners, bilinguals, and trilinguals. *Journal of Experimental Child Psychology*, *111*(3), 419–438. https://doi.org/10.1016/j.jecp.2011.09.008
Poulin-Dubois, D., Bialystok, E., Blaye, A., Polonia, A., & Yott, J. (2013). Lexical access and vocabulary development in very young bilinguals. *International Journal of Bilingualism*, *17*(1), 57–70.
R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/
Ramon-Casas, M., & Bosch, L. (2010). Are non-cognate words phonologically better specified than cognates in the early lexicon of bilingual children. *Selected Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology*, 31–36.
Ramon-Casas, M., Fennell, C. T., & Bosch, L. (2017). Minimal-pair word learning by bilingual toddlers: The catalan /e/-// contrast revisited. *Bilingualism: Language and Cognition*, *20*(3), 649–656. https://doi.org/10.1017/S1366728916001115
Ramon-Casas, M., Swingley, D., Sebastián-Gallés, N., & Bosch, L. (2009). Vowel categorization during word recognition in bilingual toddlers. *Cognitive Psychology*, *59*(1), 96–121. https://doi.org/10.1016/j.cogpsych.2009.02.002
Rocha-Hidalgo, J., & Barr, R. (2023). Defining bilingualism in infancy and toddlerhood: A scoping review. *International Journal of Bilingualism*, *27*(3), 253–274.
Samuelson, L. K. (2021). Toward a precision science of word learning: Understanding individual vocabulary pathways. *Child Development Perspectives*, *15*(2), 117–124. https://doi.org/10.1111/cdep.12408
Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, *51*(4), 1928–1941.
Schelletter, C. (2002). The effect of form similarity on bilingual children's lexical development. *Bilingualism: Language and Cognition*, *5*(2), 93–107.
Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in europe as based on levenshtein distance. *Bilingualism: Language and Cognition*, *15*(1), 157–166.
Sebastian-Galles, N., & Santolin, C. (2020). Bilingual acquisition: The early steps. *Annual Review of Developmental Psychology*, *2*(1), 47–68. https://doi.org/10.1146/annurev-devpsych-013119-023724
Sebastian-Gallés, N., Rodrguez-Fornells, A., Diego-Balaguer, R. de, & Daz, B. (2006). First-and second-language phonological representations in the mental lexicon. *Journal of Cognitive Neuroscience*, *18*(8), 1277–1291.
Shook, A., & Marian, V. (2012). Bimodal bilinguals co-activate both languages during spoken comprehension. *Cognition*, *124*(3), 314–324.

Shook, A., & Marian, V. (2013). The bilingual language interaction network for comprehension of speech*. *Bilingualism: Language and Cognition*, *16*(2), 304–324. https://doi.org/10.1017/S1366728912000466

Shook, A., & Marian, V. (2019). Covert co-activation of bilinguals' non-target language: Phonological competition from translations. *Linguistic Approaches to Bilingualism*, *9*(2), 228–252.

Singh, L. (2014). One world, two languages: Cross-language semantic priming in bilingual toddlers. *Child Development*, *85*(2), 755–766.

Smithson, L., Paradis, J., & Nicoladis, E. (2014). Bilingualism and receptive vocabulary achievement: Could sociocultural context make a difference? *Bilingualism: Language and Cognition*, *17*(4), 810–821. https://doi.org/10.1017/S1366728913000813

Spivey, M. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science*, *10*(3), 281–284.

Swingley, D. (2005). 11-month-olds' knowledge of how familiar words sound. *Developmental Science*, *8*(5), 432–443. https://doi.org/10.1111/j.1467-7687.2005.00432.x

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–166. https://doi.org/10.1016/S0010-0277(00)00081-0

Tamási, K., McKean, C., Gafos, A., Fritzsche, T., & Höhle, B. (2017). Pupillometry registers toddlers' sensitivity to degrees of mispronunciation. *Journal of Experimental Child Psychology*, *153*, 140–148. https://doi.org/10.1016/j.jecp.2016.07.014

Thierry, G., & Wu, Y. J. (2007). Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences*, *104*(30), 12530–12535. https://doi.org/10.1073/pnas.0609927104

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*(2), 172–175. https://doi.org/10.1111/1467-9280.00127

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for british english. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Von Holzen, K., Fennell, C. T., & Mani, N. (2019). The impact of cross-language phonological overlap on bilingual and monolingual toddlers' word recognition. *Bilingualism: Language and Cognition*, *22*(3), 476–499. https://doi.org/10.1017/S1366728918000597

Von Holzen, K., & Mani, N. (2012). Language nonselective lexical access in bilingual toddlers. *Journal of Experimental Child Psychology*, *113*(4), 569–586. https://doi.org/10.1016/j.jecp.2012.08.001

Wells, J. C. (1995). *Computer-coding the IPA: A proposed extension of SAMPA*. *4*(28), 1995.

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: New directions. *Annual Review of Psychology*, *66*, 173–196.

Wewalaarachchi, T. D., Wong, L. H., & Singh, L. (2017). Vowels, consonants, and lexical tones: Sensitivity to phonological variation in monolingual mandarin and bilingual english–mandarin toddlers. *Journal of Experimental Child Psychology*, *159*, 16–33. https://doi.org/10.1016/j.jecp.2017.01.009

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Zhao, X., & Li, P. (2010). Bilingual lexical interactions in an unsupervised neural network model. *International Journal of Bilingual Education and Bilingualism*, *13*(5), 505–524.

Zhao, X., & Li, P. (2013). Simulating cross-language priming with a dynamic computational model of the lexicon. *Bilingualism: Language and Cognition*, *16*(2), 288–303.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.